# UTILIZING UNTRANSCRIBED TRAINING DATA TO IMPROVE PERFORMANCE

*George Zavaliagkos, Thomas Colthurst*

BBN Technologies
GTE Internetworking
Cambridge, MA 02138

## ABSTRACT

In the past few years, the Large Vocabulary Conversational Speech Recognition (LVCSR) community has attempted to address the problem of speech recognition on languages other than English. Work on the CallHome Corpora has verified that current technology is largely language independent, and that the dominant factor with regards to performance on a certain language is the amount of training data available ([1]). This brings forth the question of what is the appropriate course of action when we need to quickly bring a recognizer up in a new language, were little or no training is available. This is exactly the question we will address in this paper. We will assume that, while only a couple of hours of transcribed data is available, much more untranscribed data can be found, and we will explore ways to utilize it.

## 1. INTRODUCTION

In the past few years, the Large Vocabulary Conversational Speech Recognition (LVCSR) community has attempted to address the problem of speech recognition on languages other than English. The data collection towards that goal was achieved by offering people free telephone calls to their native country. A number of corpora were thus created on English, Spanish, Arabic, German, Mandarin and Japanese. These corpora are generically referred to as the CallHome corpora, after the data collection method.

In a number of NIST-sponsored evaluations, it was shown that a) CallHome recognition is a very hard task, since it involves totally unconstrained conversational speech among familiar talkers (other problems include use of foreign words, high out-of-vocabulary rates, noise in the international leg of the line, simultaneous speech from more than one speaker, etc.) and b) that the technology currently used for English is very much portable to foreign languages. Word error rates vary from 50% to 60% depending on the language and the test set, and the dominant factor in determining the performance in a particular language is the amount of data collected and available for training in that language.

Since the amount of training data is so important, it is natural to ask how would one deal with the problem of building a recognizer in a new language, where none or very little training data are available. From an operational point of view, it would be ideal to only have to transcribe a few hours of speech, build a recognizer, and use this recognizer to process large quantities of untranscribed training data (which presumably is much easier to collect or have around). If the recognizer has the capability to identify the portions of the new data where automatic transcription is sufficiently accurate, we could feed that data to the training pool, thus enlarging the training set size.

The question we address in this paper is how realistic the ideal scenario presented above is. In particular, we would like to know what “sufficiently accurate” exactly means, how large does the pool of the untranscribed data need to be, and how does adding automatically transcribed (and hence corrupted data) to the training set compares with adding data transcribed by humans.

To answer all these questions we simulated the scenario presented above using the CallHome Spanish corpus. We built a model using only three hours of speech, and used it to recognize a pool of 25 hours' worth of data. We then used confidence estimation and thresholding as a means to select the “best” portion of these 25 hours, which we then added to the 3 hour training set. We found that if our confidence estimator could generate 3 more hours of speech, with the automatically obtained transcriptions that were 20% corrupted, we could obtain a small improvement by adding the new data to the training set. The improvement was bigger for new data coming from speakers seen in the training set than for speech coming from new speakers.

The paper is organized as followed: Section 2 gives a brief description of the CallHome Corpora and state-of-the-art performance across languages. It also demonstrates that the technology is largely language independent and points to the fact that the amount of training data is the dominant factor in determining performance. Section 3 explains in more detail the experimental setup for our “utilizing untranscribed data” experiment. Section 4 gives the results of our experiment, and compares the use of automatically obtained, errorful data to using human transcribed data. Finally, in section 5 we discuss the results of our experiments.

## 2. Recognition on the CallHome Corpora

In a typical recognition task in CallHome tests, we are given a (roughly) 5 minute conversation between two or more talkers. Most systems then perform a two-pass recognition; the first pass generates tentative hypotheses which are used to adapt the recognition model to each of the talkers; the second pass recognizes using the adapted models. A typical evaluation test set contains 20 such conversations. Performance on the latest NIST evaluations across languages for the BBN Byblos system are given in table 1 below

| Language | Training speech | Available text | Word Error |
|---|---|---|---|
| English | 150hrs | 3M words | 53.7% |
| Spanish | 60hrs | 0.8M words | 57.4% |
| Arabic | 18hrs | 0.3M words | 59.6% |

Table 1: CallHome recognition training data and performance across 3 languages.

As we see in table 1, there is only small difference across languages, with the error rate ranging form 53% to 60%. The high error rate is attributed to the difficulty of the tests. For example, typical OOV rates for CallHome tests is 3-4%, and for approximately 1% of the test words even the human transcribers failed to provide any transcription. Also, another comparison can be made with Switchboard [2] tests. Switchboard is similar to CallHome in that it covers spontaneous conversations, but the phone call takes place solely within the U.S and the talkers are unknown to each other. For Switchboard, the error rate is around 30%.

Looking at the amount of training data available for each of the three languages [1] we see that the amount of training data available correlates reasonably well with the performance across languages.

### 2.1. Language independence of the technology

Despite the fact that the CallHome error rates are very high, it has been demonstrated over and over again that techniques developed on other problems are also applicable to CallHome. Table 2 below gives, for example, the gain due to adaptation and Speaker Adaptive Training (SAT) [3]. It is remarkable that we get the same gain in terms of absolute reduction of the word error rate at three different operating points: approximately 5% absolute reduction in error rate, whether we start at 30% (Switchboard) or 60-70% (Arabic and Spanish).

| | Word Error % | | |
|---|---|---|---|
| | Switchboard | Spanish | Arabic |
| SI | 32.3% | 64.1% | 66.7% |
| SI-adapted | 28.2% | 61.1% | 62.6% |
| SAT-adapted | 27.2% | 59.3% | 61.5% |

Table 2: Gains due to adaptation and SAT for Switchboard and foreign CallHome

## 3. Investigating Unsupervised Training

As described in the previous section, reasonable sized corpora are available for the few CallHome languages. However, when we want to port to a new, different language *quickly*, we can only expect small amounts of training data to be transcribed and available. We would like to explore whether we can use untranscribed data (presumably available in huge quantities) to enhance the performance of models built on the minimal amounts of available training.

In particular, we will assume that

- A text corpus, not necessarily in domain, is available
- A couple of hours of speech is transcribed -preferably small amounts of data from many speakers
- Much more untranscribed data is available, and enough of it comes from the same speakers [2]. We can achieve this requirement for example by only transcribing the first few seconds of each conversation.

### 3.1. Experimental setup

We simulated the situation just described using CallHome Spanish data.

- For acoustic modeling we used 30 seconds of speech from each of 356 training speakers, for a total of 3 hours of speech.
- For language model we used transcriptions of the same data used for acoustic training (3 hours of speech, containing 42K words), augmented by 800K words of text from the ECI corpus.
- As untranscribed data, we used of speech for each of the training speakers, which gave us approximately 25 hours of speech to work with.
- As test data, we created two sets: one for speakers that were included in the training, and one for all other speakers. For the in-train set we used some of the remaining speech from the training speakers (approximately 2 hours), and for the out-of-train set we used the same test as the one used in the Fall 1996 NIST evaluation. We will refer to the two sets with the names TrainTest and Eval96.

Note that we are *simulating* the conditions presented above. In reality, we do have manual transcriptions for all these data, which we will use as a diagnostic for our work.

**System configuration:** Since we needed to process large quantities of data and our work is still in an exploratory stage, we decided to use a stripped down version of the Byblos recognizer that runs in approximately 5 times real time. For acoustic models, we trained Phonetically-Tied-Mixture (PTM) models with 64 Gaussians per mixture. Since Spanish has 36 phonemes, we used a total of 2,300 Gaussians. This is a pretty small system compared with the systems that produced the performances on table 1 (those systems use from 60K to 120K Gaussians). It should be noted however that our decision to build small models was not only constrained

[1] We should note that not all 150hrs are of English CallHome are CallHome data. The data include 134 hours of Switchboard and 16 hours of CallHome training, As we noted before, Switchboard is similar but not identical to CallHome, especially with respect to the spontaneity level of the conversation

[2] The advantage for having speech from the same speakers is that the error rate will be lower, hence we will use the new data more efficiently

by the amount of computation available, but also by the fact that we had extremely small amounts of data to train on.

For decoding we used a fast match followed by a bigram, non-crossword search, which generated n-best hypotheses for each of the test utterances. The models used to recognize the speech were *adapted* models. For each of the training speakers an adapted model was generated by supervised adaptation on the 30 seconds of transcribed speech, and the adapted model was then used to recognize the remaining untranscribed speech from the same speaker. For test speakers, the adaptation process was unsupervised: a first recognition pass was used to obtain tentative transcriptions to be used for adaptation, and a second recognition pass used the adapted model.

The n-best frequency, together with language model counts, n-gram scores and acoustic scores were input to a Generalized Linear Model (GLM) trained to generate confidence estimates for each of the words [4].

**Performance on the 3 hour training** It is interesting to compare the performance with the 3 hour training set with the performance obtained by the full Byblos system in order to identify the gains we are looking to obtain with the use of unsupervised training data.

| | TrainTest | Eval96 |
|---|---|---|
| SI | 71.9% | 78.5% |
| SI-adapted | 68.9% | 76.0% |

Table 3: Performance with the 3 hour training corpus

The performance with the 3 hour training set on both new speech form training speakers (TrainTest) and new speakers (Eval96) is summarized in table 3 above, and the comparison to the full Byblos system in table 4 below. To summarize the results, we note that, as expected, performance on training speakers (68.9% error) is better than on test speakers (76.0%), but not by a huge amount. (The difference could have been bigger if we had allocated more parameters to the training model).

Looking at table 4, we also see that for the same condition (that is no adaptation), the full Byblos model performs by 14.4% absolute better than the 3 hour model. While all this difference can be solely accounted for by the presence of the extra training, we note that if we would fix the number of parameters to the same number as the 3 hour model, the gain would only be 4.3% (from 78.5% to 74.2% word error rate). The majority of the gain (10.1%, from 74.2% to 64.1% word error rate) comes because the extra training allowed us to increase the number of parameters.

For the purposes of the unsupervised training experiment, and for the sake of simplicity, we will only try to address the effect of increasing the training size while keeping the model size fixed.

## 3.2. Confidence estimation issues

As we observe in table 3 of the previous section, the error rate of the transcription is very high, whether the test speaker was part of the training or not. Automatic transcription with the 3 hour model has an error rate of 70%-80%. Human transcribers on the other hand have an error rate of approximately 5%, and even when they make an error they usually pick a phonetically similar word. Hence, if we want to use the untranscribed training data to enhance our training set, we should be very selective in what we pick.

| | Eval96 | |
|---|---|---|
| | 3 hr, 42K word training | 58hr, 1.6M word training |
| 2K Gaussians | 78.5% | 74.2% |
| 64K Gaussians | | 64.1% |

Table 4: Comparing the 3 hour system with full Byblos (no adaptation). For both cases all of the training data were manually transcribed

A procedure for selecting the most reliable transcriptions can be based on confidence estimation and thresholding. In simpler terms, we need to have a recognizer that outputs not only the candidate word hypotheses, but also a list of its confidence that each of the hypothesized words was correct. Assuming that the output confidences correlate to true performance, we can then select a threshold on the confidence: words below this threshold are discarded, and the error rate on the retained words is controlled by the value of this threshold.

The procedure is illustrated by an example. Assume that the decoder output was the sentence

```
sentence-id    "example-utt"
hypothesis:   SIL  w1  w2  w3  w4  SIL  w5  w6  SIL
start frame:  0    10  21  42  57  63   69  81  101
confidence:       .15 .83 .91  .67       .9  .3
```

(SIL stands for silence) and that we decided that the confidence threshold was 0.8. This means that only words w2, w3 and w5 are going to be kept. For acoustic retraining, we will retain and add to the training set the following two segments:

```
    sentence-id:  "example-utt:part1"
    reference: w2 w3
    speech segment:  start 210msec; end 570msec

    sentence-id:  "example-utt:part2"
    reference: SILENCE w5
    speech segment:  start 630msec; end 810msec
```

Note that our system does not output confidence for silence frames, so we made the assumption that silence frames are going to be retained only if they are next to a word that is also retained.

For language model training we do not need to split the sentence. Instead, we can map all the low-confidence words to a "garbage" word token, so the same sentence would be added to the language model training as

```
<garbage> w2 w3 <garbage> w5 <garbage>
```

Of course, the garbage word token will not be part of the recognition lexicon.

**Selecting the confidence threshold** Table 5 presents the trade-offs between the percentage of the data that is retained and their error rate as a function of the prescribed threshold. The question of course is which threshold to select:

| threshold | % words retained | %correct in retained data |
|---|---|---|
| 0.54 | 15% | 59% |
| 0.69 | 4% | 75% |
| 0.71 | 3% | 80% |
| 0.76 | 1% | 87% |

Table 5: Trade-offs between accuracy and amount of data retained for confidence thresholding

Ideally, we would like to select the portion of the data that has the best possible accuracy. However, as table 5 indicates, for 87% accuracy we retail only 1% of the data. Since we are working with 25 hours of untranscribed data, this leaves us with only 15 minutes of data. Even if these 15 minutes were perfect, adding them to the existing 3 hour transcribed training set will have no effect. So we need to retain more.

At the 1995 Fall LVCSR workshop, Dragon systems presented an experiment where the training transcription where randomly corrupted. The baseline word error rate (no corruption) for this experiment was 55%; results presented indicated that corrupting the data by 20% caused noticeable degradation. For our experiment the baseline word error rate is 68.9%. Furthermore, the data we retain are not randomly corrupted: the recognizer hopefully picks an acoustically similar word. So we will assume that 20% error rate on the retained data is a minimum.

Unfortunately, to get 80% accuracy on the retained data, we are left with only 3% of the untranscribed data, or 45 minutes. This is still too little to have an effect when added to the original 3 hr training set. To obtain meaningful measurements we need to at least double the training size. In order to do that, we had to enhance the confidence selection. We achieved this goal by selecting a confidence threshold of 0.54, which gave us 15% of the data (3.75 hrs) at 59% correct. To obtain data with 80% accuracy we kept all the correct words (59% of 3.75hrs, or 2.21 hours) and we randomly kept only 1 out of 3 incorrect words (33% of 3.75-2.21, or 0.51 hours of speech). The overall effect is that we retained 2.72 hours of speech at 81.25% correct. Although the process we employed implies knowledge of truth, we believe that it is, in principle, fair. Had we had 100 hours of untranscribed speech, we would have been able to select a threshold of .71, to give us 3 more hrs at 80% correct, without any need for tricks.

## 3.3. Results

To summarize the experiment, we built an initial model based on 3 hours of speech ( 42K words). We then perform supervised adaptation on the 30 seconds of speech for each of the speakers, and we use these adapted models to recognize another 25 hours of speech. Based on confidence thresholding, we were able to obtain 2.72 more hours of speech (40K words), following the procedure outline on section 3.1. We added the new data to the existing training, and we retrained both acoustic and language models. The results of the new models for both training speakers (TrainTest) and new speakers (Eval96) are presented in table 6 below:

| | TrainTest | Eval96 |
|---|---|---|
| Baseline | 68.9% | 76.0% |
| Enhanced, LM only | 68.1% | |
| Enhanced, AM only | 67.7% | |
| Enhanced, AM+LM | 67.3% | 75.7 % |

Table 6: Performance on same and different speakers when we add the new data into the Language model (LM), Acoustic model (AM) , or both

As we can see, there is a small gain (0.8%) when we add the data only to the language model, a slightly bigger gain (1.2%) when we add the data only to the acoustic model, and an even bigger gain (1.6%) when we added the new data to both the acoustic and the language models. We also observe that the magnitude of the gain is much bigger for speakers that were already seen in training than for speakers that were new (1.6% versus 0.3% improvement). This is expected, since for the design of our experiment, new data from the training speakers come from the same conversation. So the data we added do not only come from the same speakers, but also from the same channel and from the same conversation topic. In a sense, the improvement for new speech from training speakers for this experiment is the upper bound on how much we could gain when we had a matched speaker, channel and topic condition. The improvement for the test speakers is a lower bound, when we have mismatched speaker, channel and topic condition. For a real application the gain will lie in between these two bounds, depending on which condition the test material reflects.

This result is encouraging. To quantify how much would the gain be had we added 100% correct data, we randomly selected another 2.72 hours from the untranscribed pool, but this time we used the human provided transcription. To minimize the randomness due to different data, we tried to make the overlap between the material selected automatically based on confidence thresholding and the new data as large as possible. The comparison between the two new sets of data is summarized in table 7 below. As we can see, the improvement for adding 20% corrupted data is roughly half of the improvement we would get if we added perfect data, for both train and test speakers.

## 3.4. Summary

In a nutshell, the conclusion of the experiment stands as follows: We with the 3 new hours having 80% accuracy in the transcription. When we fold the extra 3 hours into the training set, we observed a small improvement in performance. We observe that performance improves by half as much as

|  | TrainTest | Eval96 |
|---|---|---|
| Baseline | 68.9% | 76.0% |
| Enhanced, 80% correct | 67.3% | 75.7% |
| Enhanced, 100% correct | 65.9% | 75.4% |

Table 7: Comparison for the improvement when adding corrupted or perfect data

it would improve had we added a same amount of humanly transcribed data. The improvement is bigger for matched speaker, channel and conversation topic condition. The gain for folding the automatically generated data would have been bigger if we had increased the number of parameters of the training model.

## 4. Discussion

Although the result of our experiment is positive, one could take a negative point of view and argue that the improvement is miniscule. To see whether there is any practical implication of our experiment, we have to indulge in speculation. We will seek hypothetical answers to the following questions:

Q1 Assuming again that we start with 3 hours of speech, how much untranscribed data would we need to match the performance of systems trained on 60 hours of humanly transcribed speech?

Q2 Would the conclusions be different if the operating point were different? In other words, what if we were starting with a system whose baseline performance was 30% or less, rather than 70%?

Lets first consider question Q1. Assume that we can extrapolate our section 4 results, that is that the gain for adding 20% corrupted speech is half the gain of adding perfect data. Hence we would need 57 more hours of speech to improve the word error rate of new speakers from 76% to 67.4%. To get these 57 hours, assuming 3% retention rate, we would need 57*100/3 = 1,900 hours of untranscribed speech. And this result may be too optimistic: data that are very different from the existing model would probably perform poorly and be rejected. Hence while we add in quantity we do not diversify our training set. So to match the performance of 60 hours of humanly transcribed speech we would need *much more* than 1900 hours of speech. The amount of speech needed (two to three *orders of magnitude* more than we started with) would be probably prohibitive, and it would be faster to just annotate the speech.

However, the situation becomes more interesting when we look at different operating points. To answer question Q2, we will make the following assumption: since 20% error rate was useful for a 70% error rate system, we will assume that a ratio of 3.5 between baseline performance and corruption of the retained data is always enough. To obtain more points in the curve, we will use data from the CallHome English Spring 1997 test set and the Switchboard-II Spring 1997 test set. The error rate for these two sets with the BBN Byblos system stands at 53.7% and 35.1%, and the BBN submissions included confidence estimates. So, together with the Spanish "unsupervised training" experiment we just described, we have data points for retention based on pre-specified accuracy and confidence estimates for operational points that vary from 35% to 78%.

The result is summarized in table 8 below. As we can see, the trade-off shifts towards the automatic process; for example, for Switchboard-II we can retain 42% of the data at a 10% corruption, which may mean that we may just need closer to 10 times more untranscribed data to achieve the same effect as transcribed data.

| Corpus | W.E.R | error in retained data | retention |
|---|---|---|---|
| SWBD-II | 35.1% | 10% | 42.0% |
| CHome-Eng | 53.7% | 15% | 17.5% |
| CHome-Span.<br>3hr training | 68.9% | 20% | 3.0% |

Table 8: Retention versus corruption of retained data for various corpora

Irrespective of whether automatic training is more or less cost efficient than having human transcribers, we hope that our experiment gives a helpful insight in human learning: we just showed that even an 80% error rate system can improve itself automatically, although needing tons of data, and with the improvement coming at a turtle's pace. However, we have also demonstrated that as the system gets better, the self-learning process also accelerates, in the sense that relatively more of the new data that are encountered can be used to improve the system.

It shall also be noted that we have omitted two approaches that could improve the behavior of the unsupervised learning experiment. First, once the system improves by some measurable amount, one could iterate the process to increase the improvement. Second, the confidence estimation methods have been researched only for the past few years, so it is very plausible that better confidence estimation algorithms will become available in the future.

## 5. Acknowledgments

The authors would like to thank Manhung Siu for his help with the confidence estimation part.

## References

1. Zavaliagkos, G., et. al., "The BBN Byblos 1997 Large Vocabulary Conversational Speech Recognition System", to appear in Proc. ICASSP 98.
2. J.J. Godfrey et. al., *SWITCHBOARD: Telephone speech Corpus for research and development*, Proc. ICASSP-92, San Francisco, March 1992.
3. J. McDonough, T. Anastasakos, G. Zavaliagkos, H. Gish, *Speaker-Adapted Training on the Switchboard Corpus*, Proc. ICASSP-97, Munich, Germany, April 97.
4. M. Siu, H. Gish and F. Richardson, *Improved Estimation,Evaluation and Application of Confidence Measures for Speech Recognition*, Proc. EUROSPEECH-97, Rhodes, Greece, September 1997.